Susan Gruber*

A Causal Perspective on OSIM2 Data Generation, with Implications for Simulation Study Design and Interpretation

Abstract: Research by the Observational Medical Outcomes Partnership (OMOP) has focused on developing and evaluating strategies to exploit observational electronic data to improve post-market prescription drug surveillance. A data simulator known as OSIM2 developed by the OMOP statistical methods group has been used as a testbed for evaluating and comparing different estimation procedures for detecting adverse drug-related events from data similar to that found in electronic insurance claims data. The simulation scheme produces a longitudinal dataset with millions of observations designed to closely match marginal distributions of important covariates in a known dataset. In this paper we provide a non-parametric structural equation model for the data generating process and construct the associated directed acyclic graph (DAG) depicting the causal structure. These representations reveal key differences between simulated and real-world data, including a departure from longitudinal causal relationships, absence of (presumed) sources of bias and time ordering of covariates that conflicts with reality. The DAG also reveals the presence of unmeasured baseline confounding of the causal effect of a drug on a subsequent medical condition. Conclusions naively drawn from this simulation study could mislead an investigator trying to gain insight into estimator performance on real data. Applying causal inference tools allows us to draw more informed conclusions and suggests modifications to the simulation scheme that would more closely align simulated and real-world data.

Keywords: DAG, simulation, OMOP, NPSEM

DOI 10.1515/jci-2014-0008

1 Introduction

Prescription drugs undergo a pre-market approval process to assess safety and efficacy; however, not all drug-related adverse events (AEs) are discovered before drugs are placed on the market. Pre-market studies may lack sufficient follow-up time for detecting AEs having lengthy induction or latent periods. They are also typically underpowered for detecting rare AEs. Findings of pre-market studies may not generalize to the post-market population, who may have a higher number of co-morbidities, or be exposed through off-label applications of the drug. Since it is not possible to establish a drug's risk profile for all AEs prior to approval by a regulatory agency, there is a need for post-approval drug safety monitoring. This need motivated a consortium of pharmaceutical, FDA and academic researchers known as the Observational Medical Outcomes Partnership (OMOP), to investigate approaches to drug safety surveillance using electronic medical records (EMR) and insurance claims databases. OMOP developed the Observational Medical Dataset Simulator (OSIM2) to benchmark the performance of methods for estimating the strength of association between drug exposures and outcomes [1]. To spur innovation, OMOP sponsored the OMOP Cup, a competition for predicting adverse drug events. Algorithms entered into the competition were applied to simulated data to evaluate estimation procedures for detecting adverse drug-related events from claims data. OMOP's OSIM2 was developed to generate data similar to electronic insurance claims

*Corresponding author: Susan Gruber, Department of Epidemiology, Harvard School of Public Health, Boston, MA 02115, USA, and IMEDS Program, Reagan-Udall Foundation for the FDA, Washington, D.C., USA, E-mail: sgruber@reaganudall.org

data. A recent paper describes OSIM2 and assesses the performance of seven different approaches to analyzing the data [2]. However, that paper ignores important differences between OSIM2 data and real-world claims data that affect relative performance. This paper looks at the simulation scheme from a causal perspective to identify challenges inherent in analyzing OSIM2 data and to compare these with challenges posed by real-world claims data.

OSIM2 was designed to create a dataset that matches an actual claims dataset with respect to the number of observations and marginal distributions of key variables. However, little effort was made to capture the underlying causal structure of the observed data. In this paper we clarify the causal structure of OSIM2 data by constructing a non-parametric structural equation model (NPSEM) for the data generating process and the associated directed acyclic graph (DAG) depicting the true dependencies. (Minor variables and dependencies not germane to the discussion are omitted.) These representations reveal key differences between simulated and real-world data. Unmeasured baseline confounding is present in both, but other common sources of bias are absent from the simulated data. The longitudinal causal structure of the simulated data does not respect real-world time ordering. Although OSIM2 data provides a convenient testbed for developing tools and methodology, performance on these data may not mimic real-world performance. In this paper we apply causal inference tools to examine OSIM2. The paper is organized as follows: Section 2 introduces OSIM2 and presents an NPSEM describing the data generating procedure. Section 3 provides DAGs derived from the NPSEM and discusses their use in ascertaining identifiability of causal effects. This section also contrasts causal structure in the simulated data with that hypothesized to exist in real-world data. Section 4 demonstrates that the publicly available implementation of OSIM2 makes it difficult to accurately assess estimator bias in simulation studies. The paper concludes with a discussion of the uses and limitations of the OSIM2 simulator and offers suggestions for extensions that would bring simulated data more closely into alignment with real-world data. This work highlights the value of applying tools from causal inference to better understand the results of complex simulation studies.

2 NPSEM representation of OSIM2

OSIM2 produces a *pre-signal injection* dataset that contains no true causal drug–outcome associations and a second *post-signal injection* dataset, into which causal drug–outcome associations have been introduced. Data generation is a three-step process [3]. Step 1 consists in characterizing the distribution of the data in an observational claims dataset, with the goal of simulating a dataset that is in many ways similar to this reference data. Step 2 is the creation of the pre-signal injection dataset containing observations on each subject over time. Step 3 introduces causal drug–outcome relationships into the data to create the post-signal injection dataset. The schematic diagram in Figure 1 provides an overview of the first two steps in the process.

2.1 Step 1: fitting observational claims data

Step 1 estimates the distribution of subjects' characteristics in an external reference dataset by counting the number of subjects within strata defined by *gender*, *age*, number of medical conditions diagnosed (*conditionCount*), length of observation time (*obsTime*) and number of drug prescriptions (*drugCount*). These proportions are stored in transition probability tables consulted in Step 2.

2.2 Step 2: create pre-signal injection dataset

The goal is to generate a dataset that matches the marginal distributions of attributes in the reference data. The following NPSEM defines the dependencies that govern the joint distribution of the simulated data.

Step 1: Characterize distribution of observed claims data

Step 2: Generate observations in pre-signal injection dataset

For each subject...

- Generate baseline covariates

$P(\text{gender})$		$P(\text{age} \text{gender})$	
	Count/Total		Count/Total
M	0.65	Age	M F
F	0.35	16	0.10 0.15
Sum	1	17	0.12 0.13
		⋮	
		97	0.001 0.005
		Sum	1 1

- Add simulated medical conditions and recurrences to timeline

- Add simulated drug exposures to timeline

Figure 1: Schematic diagram of steps 1 and 2 of the OSIM2 procedure illustrating generation of the pre-signal injection dataset.

The NPSEM places no restriction on the functional forms of the relationships. Random variation is introduced through independent exogenous variables, U_x (where x indexes the variable), based on values in the transition probability tables. The first set of eqs (1) corresponds to baseline characteristics:

$$\begin{aligned}
 \text{gender} &= f_{\text{gender}}(U_{\text{gender}}), \\
 \text{age} &= f_{\text{age}}(\text{gender}, U_{\text{age}}), \\
 \text{conditionCount} &= f_{\text{conditionCount}}(\text{gender}, \text{age}, U_{\text{conditionCount}}), \\
 \text{obsTime} &= f_{\text{obsTime}}(\text{gender}, \text{age}, \text{condCountCategory}, U_{\text{obsTime}}), \\
 \text{drugCount} &= f_{\text{drugCount}}(\text{gender}, \text{ageCategory}, \text{condCountCategory}, U_{\text{drugCount}}).
 \end{aligned} \tag{1}$$

According to OSIM2 documentation, some dependencies involve only categorical versions of *age* and *conditionCount* (*ageCategory*, *condCountCategory*, respectively). The NPSEM encodes this domain knowledge in the equations for *obsTime* and *drugCount*. Substantively, this indicates that simulated data are stable with respect to slight perturbation in the distribution of the reference data.

Next, medical conditions (*condition*) are assigned to a subject's timeline in sequence, as a function of baseline covariates and the most recent previous condition (*prevCond*). The set of eqs (2) describes simulated medical conditions, how often they re-occur ($\Delta \text{days}_{\text{CC}}$), how many re-occurrences (*numEras*) and at what intervals ($\Delta \text{reoccur}$):

$$\begin{aligned}
 \text{condition} &= f_{\text{condition}}(\text{gender}, \text{age}, \text{ageCategory}, \text{condCountCategory}, \\
 &\quad \text{obsTime}, \text{time}, \text{prevCond}, U_{\text{cond}}), \\
 \Delta \text{days}_{\text{CC}} &= f_{\Delta \text{days}_{\text{CC}}}(\text{gender}, \text{age}, \text{ageCategory}, \text{condCountCategory}, \\
 &\quad \text{obsTime}, \text{time}, \text{prevCond}, U_{\Delta \text{days}_{\text{CC}}}), \\
 \text{numEras} &= f_{\text{numEras}}(\text{condition}, \text{condCountCategory}, \text{obsTime}, \\
 &\quad \text{time}, U_{\text{numEras}}), \\
 \Delta \text{reoccur} &= f_{\Delta \text{reoccur}}(\text{ageCategory}, \text{condition}, \text{obsTime}, \text{time}, U_{\Delta \text{reoccur}}).
 \end{aligned} \tag{2}$$

After all medical conditions are in place, drug exposures are added to each subject's timeline. The following eqs (3) specify non-parametric models for the number of drug exposures for each occurrence of a condition ($numDrugs_{condEra}$), the identity of the drug ($drug_{condEra}$), the number of days from onset to prescription ($\Delta days_{condEraDrug}$) and the length of the exposure ($drugDuration_{condEra}$). In these equations, $drugCountCategory$ is a categorical version of $drugCount$ and $numCurrentCond$ is a counter that is incremented as each condition is added to the timeline:

$$\begin{aligned}
 numDrugs_{condEra} &= f_{numDrugs_{condEra}}(gender, ageCategory, condCountCategory, \\
 &\quad drugCount, U_{numDrugs_{condEra}}) \\
 drug_{condEra} &= f_{drug_{condEra}}(gender, ageCategory, condCountCategory, \\
 &\quad drugCountCategory, condition, numCurrentCond, U_{drug_{condEra}}) \\
 \Delta days_{condEraDrug} &= f_{\Delta days_{condEraDrug}}(gender, ageCategory, condCountCategory, \\
 &\quad drugCountCategory, condition, numCurrentCond, U_{\Delta days_{condEraDrug}}) \\
 drugEraCount_{condEra} &= f_{drugEraCount_{condEra}}(ageCategory, condCountCategory, drug, \\
 &\quad drugCountCategory, obsTime, time, U_{drugEraCount}) \\
 drugDuration_{condEra} &= f_{drugDuration_{condEra}}(ageCategory, condCountCategory, drug, \\
 &\quad drugCountCategory, obsTime, time, U_{drugEraCount}, U_{expPerDrug_{condEra}}).
 \end{aligned} \tag{3}$$

2.3 Step 3: signal injection

Step 3 introduces causal drug–outcome relationships into the data by calculating a background level of risk and then adjusting the risk to the desired signal strength. For illustration, consider creating a causal relationship in which exposure to drug D doubles the risk of experiencing condition C within 90 days. First the background risk is evaluated by counting the number of subjects in the dataset exposed to D , n_D , and noting how many of these experience C within 90 days of the exposure, n_{DC} . The ratio, n_{DC}/n_D , represents the background level of risk. To inject a relative risk (RR) of 2, OSIM2 selects an additional n_{DC} at-risk subjects in the treatment group and inserts C into their records within the requisite 90 days. In the updated dataset, subjects exposed to D experience twice the number of events they would have experienced had D not been a cause of C . Multiple signals are injected iteratively. For each drug-condition pair, the user specifies the desired signal strength (e.g. $RR = 2$), signal type, risk window and which day of the exposure marks the start of the increase in risk. The “signal type” designation allows a risk profile to change over time (Table 1). A protective effect ($RR < 1$) is injected by deleting C from exposed subjects' records.

Table 1: OSIM2 signal-type specification defines which drug exposures are causally linked to a medical condition and the distribution of the risk over time.

Signal type	Eligible drug exposure	Selection probability*
First exposure	First	Uniform
Any exposure	Any	Uniform
Insidious	Any	Increases in proportion to days of exposure
Accumulative	Any	Accumulates over time,** ($yearsofexposure + 1$) $^\alpha$

Note: *Applies to population at risk.

**Risk increases non-linearly for $\alpha > 0$, decreases non-linearly when $\alpha < 0$ and is uniform when $\alpha = 0$.

The NPSEM equations describing medical conditions added/deleted in Step 3 ($condition_{injected}$) and the number of days after exposure that it occurs ($\Delta days_{drugCond_{injected}}$), are given in eq. (4):

$$\begin{aligned}
condition_{injected} &= f_{condition_{injected}}(drug, condition, signalType, time, obsTime, \\
&\quad U_{condition_{injected}}), \\
\Delta days_{drugCond_{injected}} &= f_{\Delta days_{drugCond_{injected}}}(signalType, riskWindow, dayExposed, \\
&\quad \alpha, U_{\Delta days_{drugCond_{injected}}})
\end{aligned} \tag{4}$$

A final, deterministic, equation defines $\widetilde{condition}$ as a composite of $condition$ and $condition_{injected}$. This equation signifies that in the post-signal injection dataset, medical conditions generated in Step 2 of the OSIM2 procedure are indistinguishable from those artificially injected in Step 3:

$$\widetilde{condition} = f_{\widetilde{condition}}(condition, condition_{injected}).$$

2.4 What we learn from the NPSEM

The NPSEM representation allows us to discover features of the data generating procedure that impact estimator performance. For example, correctly specified models for the outcome regression and the propensity score (conditional probability of drug exposure) include the total number of conditions a subject experiences and the total observation time. In reality, these variables ($conditionCount$ and $obsTime$) are summary measures that are known only at the end of follow-up. Thus we have a paradox: a parametric model-based estimator that doesn't use this information will be biased, but one that does use this information can never be applied to a real-world dataset. In this way, OSIM2 data generation violates our usual notions of causality by not respecting the real-world time ordering,

Another issue is the lack of congruence between medical conditions generated in Step 2 of the OSIM2 procedure and those artificially injected in Step 3. The NPSEM encodes the fact that the signal injection process ignores downstream longitudinal relationships: $condition_{injected}$ is a function of prior drug exposure while $condition$ is not and subsequent drug exposures and medical conditions are functions of $condition$, but not $condition_{injected}$. It's also problematic that when occurrences of a medical condition (C) are deleted to inject a protective signal, consequent drugs and medical conditions remain in the data. Because only $condition$ is recorded in the post-signal injection dataset, it is impossible to accurately model covariate–outcome relationships.

In these ways, OSIM2 favors estimators that avoid modeling exposures and outcomes. Methods that compare event rates before and after exposure are likely to perform better on OSIM2 data than methods that model the outcome regression, matching, or inverse probability weighting.

3 DAG representation

A DAG can provide a visual representation of the NPSEM. Pearl introduced a graphical way to answer questions about statistical independence known as d-separation Pearl [4]. A DAG is also useful for ascertaining identifiability of a causal association under the assumptions encoded in the DAG and for understanding how to control for confounding in order to obtain unbiased causal effect estimates.

To briefly review, variables are nodes in the DAG and causal associations are edges. Two nodes connected by a directed edge are referred to as *parent* and *child* nodes, where the parent is the source of the edge. A node's *ancestors* can be defined recursively as its parents and the parents of all nodes previously identified as ancestors. *Descendants* are defined as the node's children and the children of all nodes identified as descendants. A path between two nodes is a sequence of adjacent edges, regardless of their direction. The term *collider* refers to a node that is the child of both nodes adjacent to it along a specified path. Arrows that are present in the graph denote hypothesized causal relationships that cannot

be ruled out on the basis of prior knowledge. Knowledge of the lack of a true causal relationship between two nodes on the graph is encoded by the absence of arrows.

A path between two nodes, A and B , is said to be blocked given some set of nodes S if either there is a variable in S on the path that is not a collider for the path, or if there is a collider on the path such that neither the collider itself nor any of its descendants are in S . If all paths between A and B are simultaneously blocked by nodes in S then A and B are said to be d-separated given S and are conditionally independent given S [4]. Thus, confounding of the association between A and B can be controlled by adjusting for S in a statistical analysis.

The DAGs in Figure 2 depict the pre-signal injection dataset (only two of possibly many conditions are shown). Each panel displays all children of a highlighted node. Notably, the absence of arrows from drugs prescribed for C_1 to condition C_2 in the bottom left panel encodes the absence of causal drug–outcome associations. The post-signal injection DAG is shown in Figure 3. In contrast to the pre-signal injection DAG, $\tilde{C}_1$ and $\tilde{C}_2$ replace C_1 and C_2 . Arrows from $drug_{C1,1}$ and $drug_{C1,m}$ to $\tilde{C}_2$ correspond to drug–outcome associations injected in Step 3.

Figure 2: Each panel shows children of a specific node in the pre-signal injection OSIM2 causal DAG.

3.1 Assessing identifiability

At the top of Figure 4 is a subgraph of the post-signal injection DAG showing all edges between common ancestors of $drug_{C1,m}$ and condition $\tilde{C}_2$. The dashed line represents a causal question, *does exposure to $drug_{C1,m}$ change the risk of $\tilde{C}_2$?* Nodes that are common ancestors are circled. Variables *conditionCount* and *obsTime* are highlighted because they are not available in the dataset.

Figure 3: DAG representation of OSIM2 post-signal injection data.

Figure 4: Subgraph of the post-signal injection DAG containing all edges between ancestors of $drug_{C1,m}$ and $\tilde{C}_2$ (top). The dashed line represents the causal question of interest. Common ancestors are circled, those not recorded in the dataset are shaded. Conditioning on $\tilde{C}_1$ and prior drug exposures assigned in response to C_1 blocks confounding due to $obsTime$ (bottom).

Unbiased estimation of the causal association requires conditioning on a sufficient set of noncolliders to block all paths from $drug_{C1,m}$ to $\tilde{C}_2$. The manipulated DAG at the bottom of Figure 4 shows the effect of conditioning on the covariates in square boxes. Conditioning on $\tilde{C}_1$ blocks confounding due to $obsTime$, making it unnecessary to measure or adjust for this (unobserved) covariate. The other drugs given in response to condition C_1 potentially confound the association between $drug_{C1,m}$ and $\tilde{C}_2$. The manipulated DAG indicates that the set $\{gender, age, conditionCount, \tilde{C}_1, drug_{C1,1}, \dots, drug_{C1,m-1}\}$ d-separates $drug_{C1,m}$ and $\tilde{C}_2$.

3.2 What we learn from the DAG

Examining the DAG reveals sources of bias that must be addressed in a statistical analysis of the data. In the post-signal injection dataset unbiased estimation of the effect of $drug_{C1,m}$ on $\tilde{C}_2$ requires adjusting for

gender, *age*, *conditionCount*, $\tilde{C}_1$ and other drug exposures, $(drug_{C1,1}, \dots, drug_{C1,m-1})$. However, *conditionCount* is an unmeasured confounder, so without imposing additional modeling assumptions, the causal effect is not identifiable from the observed data. We do not need to run a simulation study to know that an analysis based on a cohort or case-control study design that adjusts only for measured confounders, will be biased. A self-controlled study design, on the other hand, inherently matches on baseline covariates, so is robust to this particular source of bias.

Suppose that instead of estimation, we were interested in classification. The OMOP Cup competition evaluated RR estimators based on their ability to distinguish between drugs that were causally associated with one or more outcome conditions and drugs that were not. For this task, cohort and case-control methods would be capable of identifying a drug that affects the risk for a particular outcome whenever (i) the magnitude of the bias is less than the signal strength or (ii) the bias moves the estimate in the same direction away from the null as the true risk.

3.3 Comparison of simulated and real-world data

Next we consider how DAGs can be used to identify other sources of bias commonly found in observational claims data in [5–7]. Figure 5 gives simple examples of structures that indicate the presence of time-dependent confounding, informative drop-out, treatment by indication and protopathic bias.

Figure 5: Subgraphs of causal DAGs illustrating time-dependent confounding (a), informative drop-out (b), treatment by indication (c) and protopathic bias (d).

Figure 5(a) illustrates time dependent confounding, where drug exposure, *drug₁* affects the value of a covariate, *bloodpressure*, that has a causal relationship with subsequent exposure, *drug₂*, and also with the outcome of interest, *C₂*. Figure 5(b) illustrates informative dropout, which occurs when a subject's probability of being lost to follow-up is a function of covariates that are also predictive of the outcome. For example, an AE (*C₂*) experienced by an employee who is too ill to work, so is no longer a member of the health care system, is not observed. Figure 5(c) illustrates treatment by indication. This occurs when drugs are more often prescribed to people perceived to be at high (or low) risk for an outcome, e.g. perceived frailty. In the example shown, a difference in the risk for *C₂* between exposed and unexposed subjects could be due to the difference in *frailty* between the two groups, rather than the exposure itself. Figure 5(d) illustrates protopathic bias. This bias is introduced when drugs are prescribed in response to early symptoms for a disease, before a definitive outcome diagnosis. For example, penicillin may be prescribed for a presumed diagnosis of strep throat (node *C₂* in the DAG), before lab results are obtained. The temporal ordering of drug exposure and diagnosis codes in the data suggests a causal drug-outcome association where none exists. In the context of drug safety surveillance protopathic bias is an important issue because it increases the number of false positives.

Upon re-examination of the post-signal injection DAG in Figure 3 we observe the presence of selection bias due to *gender* and *age*, and unmeasured baseline confounding by *conditionCount*. The DAG lacks causal structures we’d expect to see if there were time-dependent confounding, informative dropout, treatment by indication, or protopathic bias. This indicates that relative performance on OSIM2-generated data does not generalize to performance on data subject to these sources of bias. Table 2 summarizes the presence or absence in simulated data of common sources of bias in real-world data.

Table 2: Some presumed sources of bias in observational claims data.

Source of bias	Real data	OSIM2 data
Selection bias	✓	✓
Unmeasured baseline confounding	✓	✓
Time-dependent confounding	✓	✗
Informative dropout	✓	✗
Treatment by indication	✓	✗
Protopathic bias	✓	✗

4 Evaluating estimator bias

The iterative nature of OSIM2 signal injection makes it difficult to evaluate the true bias of an estimator of the RR. When multiple signals are injected one after the other, the actual signal strength will not necessarily equal the nominal signal strength. The actual signal strength depends on the degree of overlap in the exposed populations in the pre-signal injection dataset. Figure 6 illustrates this drug–interaction phenomenon when two signals are injected into the database. Nominally, $RR = 2$ for the effect of Drug A on an

Figure 6: Iterative signal injection can lead to drug-interaction effects where the true RR does not equal the nominal strength of the injected signal.

outcome (x in the Venn diagram) and $RR = 3$ for the effect of Drug B on the same outcome. The pre-signal injection dataset is shown at the top of the figure. In this dataset, 5 of the 200 subjects exposed to drug A experience the outcome event. We also see that 3 of the 120 subjects exposed to Drug B experience the outcome. Only 1 out of 40 subjects who are exposed to both drugs experiences the outcome.

The box on the left of the figure (Step 1) illustrates the injection of an RR signal of 2 for the effect of Drug A on the outcome. First the background rate of $5/200$ is calculated. Next, that rate is doubled by adding five injected conditions (x), into the data. The box on the right side of the figure (Step 2) shows what happens when injecting a signal of size $RR = 3$ for the effect of Drug B on the same outcome. The background event rate among all subjects exposed to Drug B is calculated and then this rate is tripled by adding eight new events to at-risk subjects' timelines.

In the example shown in the figure, because the populations exposed to Drug A and Drug B overlap, there is an interaction effect. Instead of exposure to Drug A doubling the risk of the outcome, the RR is 2.8. Exposure to Drug B increases the risk by a factor of 4 instead of the nominal factor of 3. Had these two populations been disjoint, there would be no drug–drug interactions and the nominal and actual signal strengths would be the same. This means that the nominal injected signal strength is not a reliably accurate reference for assessing estimator bias. Multiple signals were introduced in the publicly available OSIM2 dataset. We cannot know for certain whether interactions occurred never, seldom, rarely, or often. Therefore the bias of estimators applied to OSIM2 data cannot be assessed with confidence. This is not a desirable property of a simulator.

5 Discussion

A clear theme in talks presented at the 2013 Atlantic Causal Inference Conference session on “The Role of Causal Inference in Policy and Regulatory Decision Making” was that regulatory decision-making is orthogonal to causal inference methodology. This paper applies causal inference tools to explore the utility of OSIM2 for guiding the development of estimators that work well in practice. It demonstrates the value of incorporating a causal perspective into regulatory practice.

The NPSEM and DAG representations of OSIM2 highlight inherent barriers to obtaining unbiased causal effect estimates. Like all simulation schemes, this one does not provide a level playing field for evaluating the relative performance of different analytical approaches to estimating causal drug–outcome relationships. OSIM2 is more favorable to estimation procedures that compare pre- and post-drug exposure event rates than to those that model covariate–outcome or covariate–treatment relationships. Because of unmeasured baseline confounding, self-controlled methods will provide more robust estimates than methods that control only for measured confounders.

Relative performance of estimators on simulated data is most instructive when all the relevant challenges of real-world data analysis have been captured. In this case, many sources of bias commonly found in claims datasets are absent. Unmeasured confounding and selection bias occur in both simulated and real-world data, but the causal structures differ. Key differences include the temporal ordering of covariates and the longitudinal dependencies. We also note that in the simulated data the signal is injected without regard to biological plausibility. In reality, medical knowledge about actions within similar drug classes, biologic pathways, etc. would inform the appropriate choice of risk window, comparator group, etc. in the analysis of a single drug–outcome pair. Advantages of estimators that can exploit domain knowledge are not manifested when analyzing the simulated data.

Our findings suggest modifications to OSIM2 that would more closely align simulated and real-world data. Additional sources of bias can be mimicked by generating unobserved variables to play a role in the underlying processes. Consider an example of protopathic bias introduced when a physician prescribes penicillin for a child in advance of a definitive strep throat diagnosis. The prescription is recorded in the claims data, but the diagnosis code is added to the electronic record only after confirmatory lab results

are received. A naive analysis of the data would lead to the conclusion that penicillin increases the risk of strep. We can simulate this process by generating an unobserved covariate, *physicianBelief*, that is a function of the subject's true conditional probability of experiencing the outcome and perhaps additional sources of information already in the data. The NPSEM equations that generate *drug* and *condition_{injected}* would be modified to reflect a dependence of *drug* and *condition_{injected}* on *physicianBelief*. The magnitude of the bias is controlled by varying the importance of *physicianBelief* in generating the values of the other two variables. To mimic protopathic bias, *physicianBelief* would not be presented to the data analyst. This idea of creating latent variables to mimic sources of bias can be extended to treatment by indication and informative right censoring.

Analyzing data simulated under a broad range of scenarios helps us understand each estimator's robustness to these biases. Other existing approaches to modeling longitudinal data attempt to preserve elements of the causal structure. These include plasmoid simulation [8], simulating under marginal structural models or structural nested models with known marginal parameters [9] and generating a large dataset containing counterfactual outcomes for all treatment regimes of interest under a user-specified NPSEM.

There is value in integrating a causal inference approach with regulatory science. Because claims data are not collected for research purposes, no claims dataset will contain the information needed to accurately model every source of bias. When designing a study to address a causal question concerning a specific drug–outcome pair, information on the presence or absence of certain kinds of bias can inform the choice of analytical method. Examining OSIM2 through a causal lens can help investigators better understand the implications of OMOP simulation studies for applied work.

Funding: This work was financially supported by the U.S. Department of Health and Human Services – National Institutes of Health (grant/award number: 2 R37 AI032475-16A 1).

References

1. OMOP. Observational Medical Outcomes Partnership. 2013. Available at: <http://omop.org>.
2. Ryan P, Schuemie M. Evaluating performance of risk identification methods through a large-scale simulation of observational data. *Drug Saf* 2013;36:S171–80.
3. OMOP. Process design for the enhanced Observational Medical Dataset Simulator (OSIM 2) v1.5.005. 2011. Available at: <http://omop.org/OSIM2>.
4. Pearl J. Causality: models, reasoning, and inference. Cambridge: Cambridge University Press, 2000.
5. Daniel RM, Kenward MG, Cousens SN, De Stavola BL. Using causal diagrams to guide analysis in missing data problems. *Stat Meth Med Res* 2012;21:243–56. Available at: <http://smm.sagepub.com/content/21/3/243.abstract>.
6. Glymour MM. Using causal diagrams to understand common problems in social epidemiology. In: Oakes M, Kaufman JS, editors. *Methods in social epidemiology*. San Francisco: Jossey-Bass, 2006:Chapter 17.
7. Hernan MA, Hernandez-Diaz S, Robins JM. A structural approach to selection bias. In: *Epidemiology*. 2004;15(5):615–25.
8. Myers JA, Schneeweiss S, Rassen J. Plasmoid simulation for the evaluation of pharmacoepidemiologic methods in complex healthcare databases. Technical report, Division of Pharmacoepidemiology and Pharmacoeconomics, Harvard Medical School, 2012.
9. Young J, Hernan M, Picciotto S, Robins J. Simulation from structural survival models under complex time-varying data structures. In: *Section on Statistics in Epidemiology*, Denver, CO: JSM Proceedings, 2008.